TOWARDS AN EQUITABLE DESIGN FRAMEWORK OF DEVELOPING ARGUMENTATION IN SCIENCE ITEMS AND RUBRICS FOR MACHINE LEARNING

Paper presented at the Annual Meeting of the National Association for Research in Science Teaching (NARST). Baltimore, MD; April 2019

Tina Cheuk
Jonathan Osborne
Kathleen Remington Cunningham
Stanford University

Kevin C. Haudek
Marisol Mercado Santiago
Mark Urban-Lurain
John Merrill
Michigan State University

Christopher D. Wilson
Molly A. M. Stuhlsatz
Brian Donovan
Zoë Buck Bracey
April Gardner
BSCS Science Learning

INTRODUCTION

Argumentation is fundamental to both science and science education, to the extent that the history of science has been described as "the history of vision and argument" (Crombie, 1994, p. 3). This perspective is reflected in the Framework for K-12 Science Education (National Research Council, 2012), and the resultant Next Generation Science Standards (NGSS; Lead States, 2013) where argumentation is presented as one of eight science and engineering practices (SEPs) through which students learn the disciplinary core ideas and crosscutting concepts of science. However, it is widely acknowledged that these new standards will only have a meaningful impact if they are accompanied by high quality assessments that are closely aligned with this three-dimensional vision for teaching and learning science (NRC, 2012, Pellegrino et al., 2013). Such assessments demand a move away from reliance on the efficiency and affordability of multiple-choice items, and towards tasks that are aligned to NGSS performance expectations. In the case of argumentation in particular, the performance tasks will require students to engage in productive argumentation discourse that linguistic complex, costly, and resource intensive to score. However, efficiency and affordability remain critical components of new assessment systems, whether for research and evaluation purposes, or for broad scale state and federal measures. We therefore need new, inexpensive approaches to scoring assessments that measure three-dimensional science learning. Achieving this goal is important because "assessments operationalize constructs" (William, 2010) and if there are no assessments of argumentation that assess the performance expectations of the NGSS, it is doubtful that it will be enacted as a practice in the classroom.

Meanwhile, as educators face the emerging challenges associated with measuring these new complex constructs aligned with the NGSS, assessments at all levels are increasingly moving online. For example, PARCC and Smarter Balanced, two federally funded testing consortia, both use computer-based assessment systems. There are also a number of research groups exploring how simulations and digital learning environments can be used to measure three-dimensional learning (NRC, 2014). However, such assessments are still limited by their inability to score student written work efficiently. Further, if the writing of students during online learning experiences could be scored in real time, both formative information could be supplied to the students and their teachers, potentially influence classroom instruction.

Recognizing that designated English Learners (ELs) make up nearly 10% of our student populations in public schools, assessment developers need to consider how language and cultural perspectives may influence how students interpret and respond to science items. These include "the values, beliefs, experiences, communication patterns, teaching and learning styles, and epistemologies inherent in students' cultural backgrounds, as well as the socioeconomic conditions prevailing in their cultural groups" (Solano-Flores & Nelson, 2001, p. 553). How we assess across cultures often relies on implicit or tacit cultural assumptions about student populations. If the assessments are designed for the dominant culture of English-only students, then the inferences we are making about ELs may be biased and underestimate their ability.

For all students, the practice of argumentation in science will test both their receptive and productive English literate practices as they engage in sense-making with the item and produce textual—in this case typed English responses—into a computer system. At the same time, all students must navigate the science knowledge, procedural knowledge and epistemic knowledge that are inherent within the practice of scientific argumentation. As a result, we have developed a set of item *design principles* as part of an emerging design framework that incorporates these considerations into our research, design, and development (RDD) cycle. Guided by these principles, our goal is to develop accurate and reliable scoring models that are able to score students' written responses at levels equal to human expert scorers, and to accurately place students on a learning progression for argumentation.

This paper draws from work of a larger RDD project where we used automated lexical analysis and machine learning techniques to develop valid and reliable constructed response measures of student scientific argumentation that can be administered and scored at scale. Our primary research question for the project was as follows:

> *How can automated lexical analysis and machine learning techniques be applied to developing an efficient, valid, and reliable measure of students' placement on a learning progression for argumentation?*

This design framework draws from the iterative process of question design, collection of data, and rubric development, bringing together the domains of argumentation, learning progression, and automated analysis.

METHOD

Our interdisciplinary team members, composed of experts in argumentation, assessment development, and language development, developed three sets of middle school science items, each in a different disciplinary context, and rubrics intended to assess argumentation in science to be scored by both human and machine learning models. In our RDD work, we utilized the crowdsourcing platform Amazon Mechanical Turk (MTurk) to recruit U.S. adults (n= 100) in our pilot phase, 246 8th grade students from a mid-sized urban school district in Northern California, in our beta phase. Our final data collection came from two sources: A private independent school (grades 5-8) with approximately 100 student responses and a set of five middle school (grades 6-8) science classrooms with approximately 900 student responses from a public school district in the California Bay Area, totaling about 1000 responses that were then coded by experts and used in the machine learning modeling and analysis.

In each of these phases, our team made revisions and improvements to the items and documented the changes we made throughout the process. It was in this testing and revision process that these principles emerged from our work in the past eighteen months of development work.

AN EMERGENT DESIGN FRAMEWORK

These emergent design principles have guided both item and rubric development, especially pertaining to scoring items that assess argumentation in science, initially scored by humans which in turn informs machine learning scoring. We are guided by the Question Development Cycle (QDC) workflow in Figure 1 and focused the design and development work at the question design and rubric development steps while using student responses and human coding to iterate on both question design and rubric development.

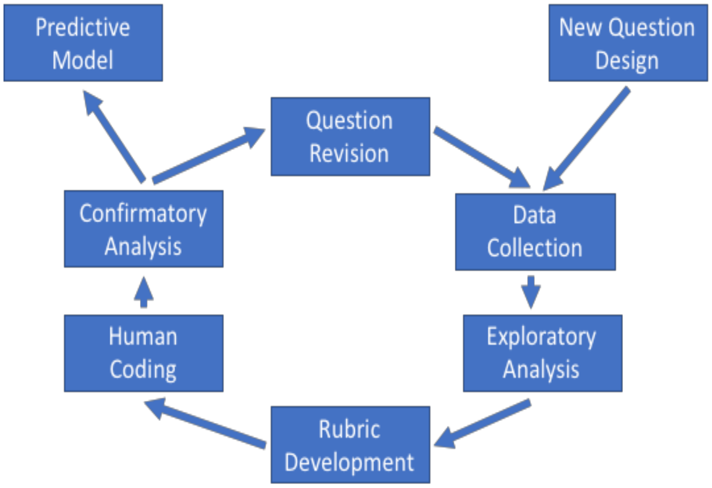


Figure 1. Overview of the Question Development Cycle (QDC) workflow.

We organize this design framework across three principles:

Principle 1: Designing for argumentation in science

Principle 2: Designing for students' language interactions
Principle 3: Designing for machine learning

Within each of these dimensions, we discuss important principles that emerged through the iterative design and development processes, undergirded by a goal of equity for students who may be designated ELs. Using illustrative examples, we highlight how these principles are enacted and operationalized so that these principles could be transferrable and adapted in similar RDD contexts.

CORE PRINCIPLE 1: Designing for argumentation in science

In their research to date, Osborne and his team have developed a learning progression for argumentation in the context of the structure of matter (Osborne, Henderson, MacPherson, & Szu, 2016). The construct map in Table 1 provides a summary of the learning progression.

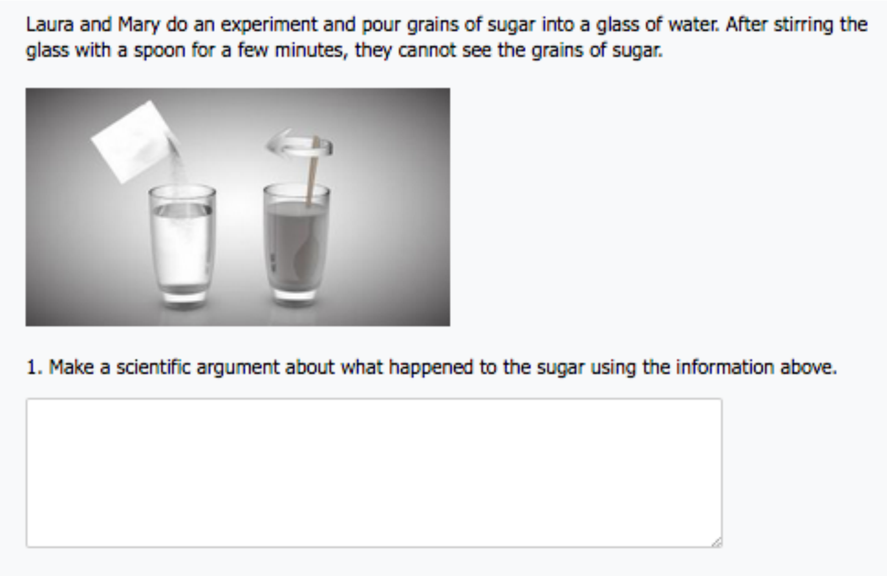
Level	Constructing	Critiquing	Description
0a	Constructing a claim		Student states a relevant claim.
0b		Identifying a claim	Student identifies another person's claim.
0c	Providing evidence		Student supports a claim with a piece of evidence.
0d		Identifying evidence	
1a	Constructing a warrant		Student constructs an explicit warrant that links their claim to evidence.
1b		Identifying a warrant	Student identifies the warrant provided by another person.
1c	Constructing a complete argument		Student constructs a synthesis between the claim and the warrant.
1d	Providing an alternative counter argument		Student offers a counterargument as a way of rebutting another person's claim.
2a	Providing a counter-critique		Student critiques another's argument.
2b	Constructing a one-sided comparative argument		Student makes an evaluative judgment about the merits of two competing arguments
2c	Providing a two-sided comparative argument		Student provides an evaluative judgement about two competing arguments
2d	Constructing a counter claim with justification		Student explicitly compares and contrasts two competing arguments, and an argument as to why it is superior to each of the previous arguments.

Table 1. Scientific Argumentation Construct Map, (Osborne et al., 2016)

The map for argumentation is innovative in that it includes critique, which is essential for scientific argumentation as the construction of knowledge is a dialectic between construction and critique (Ford, 2008). In other words, being able to explain why an idea is flawed is as important as being able to explain why it is right. Empirical work to date has shown the construct to be psychometrically uni-dimensional, and that it supports the distinction between the two columns (Constructing vs. Critiquing) in that the rows shown in Table 1 have different difficulties (Yao, 2013). Like all construct maps, it defines a continuum of understandings, providing a "coherent and substantive definition for the content of the construct" (Wilson, 2005). In other words, level 0 items are easier than level 1 items, and level 2 items are more difficult than level 1, and 0 items respectively. Within each level, the letters “a, b, c, d” denote similar levels of difficulties *within* a particular level, and do not necessarily signal a progression of understanding within the level.

Sub-Principle 1a. Alignment across question design, expected student responses, and rubric development to a specific level on the learning progression (LP).

In the design of each item within a set, we targeted *one* of the 12 possible levels that have been identified in Table 1. As seen in Figure 2, we designed an item to assess LP “1c” constructing a complete argument. The question stem of “Make a scientific argument about what happened to the sugar using the information above” aims to assess students’ abilities in constructing a complete argument.



Laura and Mary do an experiment and pour grains of sugar into a glass of water. After stirring the glass with a spoon for a few minutes, they cannot see the grains of sugar.

1. Make a scientific argument about what happened to the sugar using the information above.

Figure 2. Sample sugar item at LP “1c”.

Sub-Principle 1b-1. Accept the “2-NGSS-dimensionality” of items that assess *both* argumentation practice (SEP) and disciplinary core ideas (DCI).

Students’ argumentation practice in science is a complex construct in that students are tasked to engage in the argumentation *in* a specific science discipline or DCI. While the LP from Osborne and his team signaled that it would be a uni-dimensional construct, in our development work, we acknowledged that students will have to understand and interpret the science content that is inherent in the task *and* be able to produce scientifically-accurate science knowledge in the discourse of argumentation in their response. In other words, in our design of the items and rubrics, we did not isolate argumentation practice away from the science content and made a deliberate design decision to consider *both* NGSS dimensions, SEP and DCI, into assessment design.

Sub-Principle 1b-2. Minimize prior science content knowledge students would need to interpret and respond to task if the *primary goal* is to assess argumentation.

Even though we accept the "2-NGSS-dimensionality" of our items, our team had a primary goal of using the LP of argumentation that foregrounded our item design. As a result, we designed items to minimize the prior science knowledge student would need to interpret and respond to the task so the primary construct of interest would be students' argumentation practice. This design element was further reflected in the rubric development in that we first prioritize how students would engage in argumentation practice, then we considered how students communicated their science knowledge *through* the argumentation components. In other words, our primary focus was on student's facilities with argumentation, given the context of the item, instead of primarily scoring for science content knowledge.

Sub-Principle 1c. Provide an appropriate level of scaffolding for tasks and recognize the cognitive load appropriate for age group based on the difficulty of the item (Mislevy & Duran, 2014). This allows us to assess a specific level of the learning progression.

Building on prior principles 1a, and 1b-1 and 1b-2, we scaffolded our items by asking easier leveled items from the LP first, then followed by more difficult items at higher levels. Rather than start the item set with a question at level 2, we typically started our item set at level 0 or 1, and moved progressively to higher levels. This progression of levels was intended to reduce the cognitive load for students by creating what we called a "low-floor" entry point in accessing the item and thereby building item difficulty into subsequent items.

CORE PRINCIPLE 2: Designing for students' language interactions (receptive and productive functions), with the task and their expected written outcomes.

Our work in this principle on language interactions is focused on equity-based outcomes for designated ELs that are a growing sub-population in the U.S. public schools. We considered how students might make sense of the language demands of the tasks in their receptive and productive uses when engaged in argumentation with evidence (Solano-Flores, Barnett-Clarke, & Kachchaf, 2013). Receptive uses of language are difficult to observe as these uses happen when students reading, interpreting, and understanding the task at hand. Productive uses of language are observable and measurable, and are seen in students' production of language through written and spoken modalities. Students are using language towards *functional* purposes in communicating meaning within the context of science learning (Halliday & Martin, 1993). At the same time, the *structural* components of argumentation undergird *how* students interpret and communicate their understanding of the task situation.

It is through students' productive language use that we are able to measure and thereby understand what students know and can do in the context of a DCI. This role of language and how we think about it in assessment terms with machine learning drives our design and development processes. In other words, we wanted to ensure that our task assesses the constructs that we set out to assess. To do that, we considered the receptive and productive language demands that students would need to engage in within the "2-NGSS-dimensional" model (see principle 1b-1) in the forms of argumentation practice and science knowledge (DCI).

We operationalized principle 2 by creating a "language audit" composed of the following four questions, subdivided by receptive and productive language demands as shown in Figure 3.

Receptive language demands:

1. What science constructs / disciplinary core ideas (DCI) do students need to interpret and reason within the task?
2. What elements (or components) of argumentation practice must students understand within the task?

Productive language demands:

3. What science constructs must students demonstrate knowledge in?
4. What elements (or components) of argumentation practice must students demonstrate competency in? [Reflective of the LP level assessed]

Figure 3. Language audit of the task

This particular step is important in how we think about designing rubrics for machine learning. Because machine learning is dependent on students' linguistic output, our team had to think thoroughly about how students would demonstrate their facilities with argumentation within a science task scenario.

Additional language specific considerations that we incorporated into the item development, and consequently rubric development included:

- Maximize construct-relevant language and minimize construct-irrelevant language (Abedi, 2004; Solano-Flores, 2006).
- Consider how the various multi-modal representations that are distinctly found in science classrooms may augment or hinder students' sensemaking and ultimately their productive language functions in responding to the task (Cook, 2006; Lemke, 1998; Solano-Flores et al., 2014).
- Be inclusive of students' science "register". Accept how students use their home language or everyday language (e.g., words, analogies, examples) to show how they are making connections to science concepts and how those registers could be used to engage in argumentation discourse.
- Consider how pronouns are used in reference to subjects and objects both within the item and how students may use pronouns in their written responses.
- Sample represents and is inclusive of diverse linguistic repertoires.

CORE PRINCIPLE 3: Designing for machine learning

A supervised machine learning text classification approach assigns student written responses a score using an algorithm that has been trained on a set of known data (see Aggarwal & Zhai, 2012). In order to create the appropriate input from student responses for the machine learning model, we had to design items and respective rubrics that allows us to isolate discrete linguistic components of student responses. In addition, since this classification approach "learns" from human coded data, we had to be certain that the human codes were reliable across items and contexts and focused on the critical features of argumentation as identified in the learning progression.

Sub-Principle 3a. Rubric is designed to capture the "analytical parts" that make up the whole. These "analytical parts" are then used to create the holistic models in scoring. The design of the item and rubrics allows us to isolate discrete linguistic components in student responses.

Analytical components need to be manageable for human scoring. Student responses along with consensus scores done by humans are input into a machine learning application to generate a scoring model. If the resulting model has an appropriate inter-rater reliability (with a desired Cohen's kappa of 0.8 or higher) between human and machine scores, we generate a final predictive model to score future student responses.

Sub-Principle 3b. Analytical component is scored with a binary, *a priori* approach (0/1). A binary approach (0/1) allows us to focus on the presence or absence of students' ideas relevant to argumentation in science.

Because supervised machine learning is dependent on human scored responses and detectable word patterns in a sample, we designed our rubrics to be relatively easy to score with discrete analytical components for human scorers. The design of the rubric allows us to identify the analytical components that student responses would need in order to contribute to a holistic score. It is important to note that these analytical components represent students' thinking at a fairly fine grain size that are inclusive of the structural components LP of argumentation (e.g., claim, reasoning, evidence, etc.) *and* within the DCI that is assessed (see principle 1b-1 "2-dimensionality NGSS of items").

Sub-Principle 3c. Design goal is to target the right level of difficulty so that there is variation in scoring (both at the component and holistic level).

Lastly, in our design of items, we had to think about targeting the right level of item difficulty so that the scoring at both the component and holistic levels would have enough variation to "train" the machine learning model. In other words, having a set of components that are scored as all "0" or all "1" is not as useful to building a machine learning as compared to a mix of "0" and "1". Both the design of an item and its rubric have direct implications to how humans score student responses to identify analytical components in them. All the considerations in principle 2 may contribute to the diversity of possible student responses, which in turn influence the range of analytical components that are part of the rubric, and subsequently scored by humans. The greater linguistic variation we have through sampling and thoughtful analysis of the receptive and productive language resources that students are using, the better we can design assessments that are inclusive and equitable for designated ELs in our school systems.

CONCLUSION

Educational reforms demand assessments move away from reliance on the efficiency and affordability of multiple-choice items, and towards the use of more authentic tasks aligned to broader skills and performance expectations. We therefore need new, less time-intensive approaches to scoring assessments, while guided by principles that create equitable assessments for linguistically diverse learners. The framework described here addresses specific design issues that emerged from our development of items and rubrics in measuring students' ability to engage in scientific argumentation.

Our contribution to the field is grounded in students' linguistic resources they bring to the task as they engage with and respond to the task in the norms of argumentation discourse. These principles thereby *foreground* much of the work done in the assessments for diverse learners in that we design with student's linguistics resources and the language demands of the task at the *center* and subsequent steps of the QDC workflow (Figure 1), not as an afterthought in modifying the conditions by which students might be engaged with the task (e.g., provided more time, use of a glossary, simplifying task, translating task in home language). The through-line for this work is about language and ways students use language to perform argumentation discourse in science. Machine learning, therefore, serves as mechanism for us to analyze this rich and complex written discourse, potentially informing how and what students know and can do, and influence instruction.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. DUE 1561159. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Aggarwal C.C., Zhai C. (2012) A Survey of Text Classification Algorithms. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA

Cavagnetto, A., Hand, B. M., & Norton-Meier, L. (2010). The Nature of Elementary Student Science Discourse in the Context of the Science Writing Heuristic Approach. *International Journal of Science Education, 32*(4), 427 - 449.

Crombie, A. C., (1994). *Styles of scientific thinking in the European tradition: The history of argument and explanation especially in the mathematical and biomedical sciences and arts (Vol. 3)*. London: Duckworth.

Ford, M. J. (2008). Disciplinary authority and accountability in scientific practice and learning. *Science Education, 92*(3), 404-423.

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: prospects and limitations. *CBE—Life Sciences Education, 10*(4), 379-393.

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid–base chemistry in introductory biology. *CBE—Life Sciences Education, 11*(3), 283-293.

Jurka,TP., Collingwood, L., Boydstun, AE., Grossman, E., Van Atteveldt, W. (2012). RTextTools: Automatic Text Classification via Supervised Learning. R package version 1.3.9. http://CRAN.R-project.org/package=RTextTools

Kelly, G., & Takao, A. (2002). Epistemic Levels in Argument: An Analysis of University Oceanography Students' Use of Evidence in Writing. *Science Education, 86*, 314-342.

Lee, O., Quinn, H., & Valdés, G. (2013). Science and Language for English Language Learners in Relation to Next Generation Science Standards and with Implications for Common

Core State Standards for English Language Arts and Mathematics. *Educational Researcher, 42*(4), 223-233

Lee, H.-S., Liu, O. L., Pallant, A., Roohr, K. C., Pryputniewicz, S., & Buck, Z. E. (2014). Assessment of uncertainty-infused scientific argumentation. *Journal of Research in Science Teaching*, *51*(5), 581–605. https://doi.org/10.1002/tea.21147.

Liu, O. L., Brew, C., Blackmore, J., & Gerard, L. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement-Issues and Practices*, *33*(2), 19–28. https://doi.org/10.1111/emip.12028

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H.-S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, *23*(2), 121–138. https://doi.org/10.1080/10627197.2018.1427570

Mislevy, R. J., & Duran, R. P. (2014). A sociocognitive perspective on assessing EL students in the age of Common Core and Next Generation Science Standards. *48*(3), 560-585.

Moharreri, K., Ha, M., & Nehm, R. H. (2014). EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach, 7*(1), 1-14. doi:10.1186/s12052-014-0015-2

National Research Council. (2007). *Taking Science to School: Learning and Teaching in Grades K*-8. Washington DC: National Research Council.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, D.C.: The National Academies Press.

National Research Council (2014). Developing assessments for the Next Generation Science Standards. Washington, DC: National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: for States, By States*. Washington, DC: The National Academies Press.

Osborne, J. F., Henderson, B., MacPherson, A., Szu, E., Wild, A., & Shi-Ying, Y. (2016). The Development and Validation of a Learning Progression for Argumentation in Science. *Journal of Research in Science Teaching, 53*(6), 821-846.

Osborne, J. F. (2010). Arguing to Learn in Science: The Role of Collaborative, Critical Discourse. *Science, 328*, 463-466.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (Eds.). (2013). *Developing Assessments for the Next Generation Science Standards*. Washington DC: National Academies Press.

Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. *Journal of the Learning Sciences, 12*(1), 5-51.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*(5), 553–573.

Solano-Flores, G., Barnett-Clarke, C., & Kachchaf, R. R. (2013). Semiotic structure and meaning making: The performance of English language learners on mathematics tests. *Educational Assessment*, *18*(3), 147-161.

Toulmin, S. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.

Urban-Lurain, M., Prevost, L., Haudek, K. C., Henry, E. N., Berry, M., & Merrill, J. E. (2013). *Using computerized lexical analysis of student writing to support Just-in-Time teaching in large enrollment STEM courses.* Paper presented at the 2013 IEEE Frontiers in Education Conference (FIE).

Yao, X. M. (2013, March). Automated Essay Scoring: A Comparative Study. In Applied Mechanics and Materials (Vol. 274, pp. 650-653).

William, D. (2010). What Counts as Evidence of Educational Achievement? The Role of Constructs in the Pursuit of Equity in Assessment. *Review of Research in Education, 34*, 254-284.

Wilson, M. (2005). *Constructing Measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.